TEC-0041

AD-A276 658

Selection of Climate Station Data Using Clustering and Triangulated Irregular Network Techniques

Kevin R. Slocum

DTIC ELECTE MAR 08 1994 S B D

October 1993

DTIC QUALITY INSPECTED 3

Approved for public release; distribution is unlimited.

U.S. Army Corps of Engineers
Topographic Engineering Center
Fort Belvoir, Virginia 22060-5546

US Army Corps of Engineers
Topographic Engineering Center

TEC

94-07513

Best
Available
Copy

Destroy this report when no longer needed.
Do not return it to the originator.

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

The citation in this report of trade names of commercially available products does not constitute official endorsement or approval of the use of such products.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
October 1993

3. REPORT TYPE AND DATES COVERED
Technical Report Mar. 1992 - Jan. 1993

4. TITLE AND SUBTITLE
Selection of Climate Station Data Using Clustering and Triangulated Irregular Network Techniques

5. FUNDING NUMBERS
4A162784A855
JO, 401

6. AUTHOR(S)
Kevin R. Slocum

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Topographic Engineering Center
Fort Belvoir, VA 22060-5546

8. PERFORMING ORGANIZATION REPORT NUMBER
TEC-0041

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of the Chief of Engineers
Washington, DC 20314-1000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This report presents a methodology of adding an elevation variable to the process of selecting climate station data for computer modeling. Various techniques are used in this method which include statistical correlation of climatic variables, clustering of climate station data, and development of a triangulated irregular network (TIN) internal graphic layer with nodes representing climate station locations. Improvements to the climate station data selection process are documented.

14. SUBJECT TERMS
Correlation, clustering, triangulated irregular networks, climatology, modeling

15. NUMBER OF PAGES
32

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

TABLE OF CONTENTS

Accession For
NTIS GRA&I ☑
DTIC TAB ☐
Unannounced ☐
Justification
By
Distribution/
Availability Codes
Dist | Avail and/or Special
A-1

ILLUSTRATIONS

TABLES

FIGURES

PREFACE

This report was conducted under DA Project 4A162784A855, JO, 401, "Environmental Decision Support". The study was conducted during March 1992 through January 1993 under the supervision of Mr. Jerry M. Breen, Chief, Terrain Support Branch, and Mr. Paul G. Bourget, Chief, Environmental Effects Branch, and Mr. Bruce K. Opitz, Director, Geographic Sciences Laboratory, U.S. Army Topographic Engineering Center (TEC), Fort Belvoir, Virginia 22060-5546.

The author wishes to express his appreciation to Mr. Paul F. Krause of the Geographic Sciences Laboratory, (TEC), for his numerous contributions in defining climatological considerations critical to this project. The author also acknowledges the efforts provided by both Mr. Vinh Duong of TEC and Dr. Christoph Witzgall of the National Institute of Standards and Technology for their important roles in software programming support.

Mr. Walter E. Boge was Director, and Lieutenant Colonel Louis DeSanzo was Commander and Deputy Director of the U.S. Army Topographic Engineering Center at the time of publication of this report.

SELECTION OF CLIMATE STATION DATA USING CLUSTERING AND TRIANGULATED IRREGULAR NETWORK TECHNIQUES

INTRODUCTION

Purpose

This research was conducted to develop and implement an improved methodology for selecting climate stations to represent areas of the world where climate data is unavailable. Selection of an incorrect station would result in the return of climatic data that was not indicative of the geographic area of interest. Any terrain modeling results that used incorrect climatic data input would subsequently be of dubious value.

The methodology detailed in this research paper was tested and evaluated against the present technique for climate station selection. In a sampling of 30 locations within Germany, the newly developed methodology returned a more reliable climate station selection for 10 of the locations.

Background

Two climate station selection techniques have routinely been utilized within the US Army Corps of Engineers (USACE) computer models for accessing the most appropriate climate station data. Both techniques are inadequate for accurately selecting climate stations, however.

The first technique for climate station selection relied on the identification of all worldwide climate stations as members of geographical location categories, such as, (a) coastal, (b) inland, or (c) mountain. A computer terrain model would then define an 'observer location' as either coastal, inland, or mountain category. Selection was based on proximity to the nearest climate station that was within a matching category. Several problems existed with this approach.

- Groupings were subjectively arrived at and could not be replicated.
- Knowledge of observer location category was necessary, but not necessarily known.
- Designation of the wrong grouping guaranteed that a correct climate station and its data would never be recalled from the data base.

The second technique for climate station selection simply selects the closest climate station to a user-defined observer location. This technique abandoned the arbitrary category method previously used and instead included distance from stations to observer location as the only consideration. This exclusive use of distance ignored other factors affecting climatology at a location such as latitude, proximity and relationships to such features as water bodies and mountain barriers, local topography, elevation, and dominant air mass controls.

Enhancing the climate station selection process to include, at least initially, one of the many missing determinants of a climate was an immediate goal. Elevation was selected as the first additional variable as it was deemed to be the most important. Digital Terrain Elevation Data (DTED) Level 1 from the Defense Mapping Agency (DMA) was used. It was incorporated into the climate station selection process by ensuring that observer location elevation and climate station elevation had similar heights. Additional variables were considered, but resources were limited and inclusion of more than one variable would have compounded the time and effort required. Also, elevation source data was the only variable readily available. Climate factors other than elevation were regrettably omitted for this initial research, but they are planned for future iterations into this subject.

Intended Application of Models. Climate data are intended to be used for modeling during times when real-time meteorological information is either unavailable or inappropriate. For example, long-range planning missions would not rely on meteorological data but on historical information contained in a climatological data base. For near-term mission planning, the absence of meteorological data would necessitate the use of climate data as an alternative source of information. Data sparse areas are not characterized by data derived through interpolation from known climate stations but, instead, are emulating the data of the most appropriate climate station. A first step to determine a most appropriate station is the issue addressed in this report. Interpolation of climate data is intended for future follow-on research.

Literature Review. Climatologists have geographically modeled various climatic parameters over the years, and the resulting mapped products are typically the result of interpolation between climate station data points.[1] For example, Hutchinson seeks to estimate a rainfall surface using irregularly spaced and weighted climate station points.[2] A wealth of climatology source material exists that examines the influences and implications of temporal climatological change.

This study is a blend of statistical applications (correlation and clustering) and is used in conjunction with a triangulated irregular network (TIN) borrowed from the topographic sciences. The statistical application usually least understood is clustering. The goal of cluster analysis is to detect interrelationships and like characteristics between the data describing some cases (in this instance, stations) and then to place these cases into relatively homogeneous groups.[3] Anderberg provides an excellent review of clustering and how it can be applied to real-world scenarios.[4] Although numerous other references were available, they were not as appropriate to this research project. Clustering was also well documented within the chosen statistical software package.[5]

Use of triangulated irregular network (TIN) software is well documented in many published sources. The predominant discussion is TIN capability within the digital elevation model arena. The digital elevation models are then used for applications such as slope, viewsheds, and aspects.[6] More

[1] See for example, H.Landsberg, Physical Climatology, Gray Printing Co., Inc., DuBois, Pennsylvania, 1962.

[2] M.F. Hutchinson and R.J. Bischof, "A New Method for Estimating the Spatial Distribution of Mean Seasonal and Annual Rainfall Applied to the Hunter Valley, New South Wales," Australian Meteorological Magazine, No. 31 (1983), p. 179-184.

[3] Norusis/SPSS Inc., SPSS-X Introductory Statistics Guide, 1988.

[4] M.R. Anderberg, Cluster Analysis for Applications, Air Force Systems Command, United States Air Force, Academic Press, Inc., New York, New York, 1973.

[5] D. Wishart, Clustan User Manual, 4th Ed., Computing Laboratory, University of St. Andrews, 1987.

[6] See for example R.A. Pries and R.A. Schowengerdt, Computer Assisted GIS Data Entry at the Bonneville Power Administration, GIS/LIS '88, American Congress on Surveying and Mapping, 1988, p.1-10.

obscure applications do exist, however, as where TIN's have been used for archaeological site selections and even fossil pollen identification.[7]

The U.S. Army Topographic Engineering Center (USATEC) retains a voluminous amount of historical climate data. Some of these data are stored digitally in a program called the Battlefield Environmental Effects Software (BEES).[8] At present, a total of 605 worldwide climate stations are stored in the data base with 138 possible climate parameters. Climatic elements contained in this digital data base encompass various treatments of temperature, precipitation, humidity, ceiling heights, cloud cover, visibility, wind, and atmospheric pressure. Any one station retains monthly data for up to 36 different climate parameters. Stations do not collect information on identical climate parameters. A 43 station subset of the 605 possible stations were selected for this research project. All 43 stations are located in Germany.

A total of 605 climate stations worldwide creates an extremely sparse network. A large expanse of land may exist for which climatic characteristics do not exist. Given a geographic coordinate and elevation for a position on the ground, an assessment can be made as to which climate station would best exemplify the climatic conditions expected at that location on the earth.

The AirLand Battlefield Environment (ALBE) is a suite of application software that demonstrates tactical decision support aids to assist in military mission planning.[9] The BEES climate data is incorporated into several of the ALBE software routines and contributes to the overall outcome of these products. Knowledge of terrain and weather conditions are critical factors for successful exploitation of the battlefield environment. Accordingly, incorrect selection of climatic stations implies incorrect climate data selection. The weather component of the battlefield has not been omitted in this case, but worse yet, is erroneously conveyed.

METHODOLOGY

Research Steps

Methodology described in this report details an improved technique for providing reliable climate station selection. Naturally, as the station selection process improves, so too does the corresponding climate data which is critical to accurate terrain modeling purposes. Research was conducted in four distinct steps:

---

[7] S.E. Howe, Estimating Regions and Clustering Spatial Data: Analysis and Implementation of Methods Using the Voronoi Diagram, Division of Applied Mathematics, Brown University, PhD Dissertation, October 1978.

[8] BEES is a software product distributed by the U.S. Army Topographic Engineering Center, ATTN: CETEC-GL, Fort Belvoir, VA 22060-5546, telephone (703) 355-2840. Information pertinent to this product should be addressed to the Environmental Sciences Division.

[9] AirLand Battlefield Environment (ALBE) is a prototype testbed for demonstrating tactical decision aids. Information related to the ALBE program can be obtained by contacting: U.S. Army Topographic Engineering Center, ATTN: CETEC-GL (J. Breen), Fort Belvoir, VA 22060-5546, telephone (703) 355-2855.

1. **Correlation:** The correlation package selected the 'candidate' climate variables to be used for clustering. A final candidate selection list and rationale for variables to be used were created.

2. **Clustering of Climate Stations:** Clustering used the correlation variables, manipulated them through a homogenous grouping algorithm, and assigned climate station cluster groups.

3. **Selection of a Point-in-Polygon Method:** Thiessen polygons, manually produced polygons, and triangulated irregular networks (TIN's) were each considered. The TIN's were chosen because they provided a reliable mechanism for examining distance and elevation factors. The TIN provided spatial intelligence within the model and identified instances when an observer location was located between defined climate zones.

4. **Test and Evaluation:** The results of this new methodology for climate station selectior. were tested and evaluated against the distance criteria only method for station selection. The categorical station selection method was dismissed as being outdated and unnecessary for comparison.

A discussion of each of these research steps follows.

Correlation

Climate variables for all 43 German climate stations were correlated against one another in an effort to minimize the number of variables needed for later statistical clustering.[10] Possible variables for any station were reduced from 36 per month to a maximum of 10 climate variables per month. Dilution of the variables was necessary to satisfy size constraints placed by the statistical software programs used. Only variables reported by over 95 percent of the stations were considered from the correlation process. This 95 percent threshold minimized the potential problems of correlating against fields of information for which missing data was prevalent. Table 1 presents the inclusive list of variables that were initially considered candidates for correlation.

Several of the following 16 variables were eliminated from consideration in the correlations by in-house climatological experts. Variables 6, 7, 12, 13, 14, and 15 were eliminated because they were viewed as being of trivial importance compared to the variables to be retained for final correlation. The correlation was measured for the degree of association between the different monthly climate parameters of the 43 climate stations. Temperature variables 1-5 (measured in degrees Fahrenheit) were compared against each other while precipitation variables 8-12 (measured in inches) were compared against each other. Station elevation was added to the list of variables as a non-climatic variable.

[10] The PC based clustering routines to be used had a limited memory resource. Only a limited number of variables and data points could be accommodated.

Table 1. Candidate correlation variables

Variable 1 -	absolute max temperature
Variable 2 -	average daily max temperature
Variable 3 -	average monthly temperature
Variable 4 -	average daily min temperature
Variable 5 -	absolute min temperature
Variable 6 -	average number of days w/temperature < =32
Variable 7 -	average number of days w/temperature < =0
Variable 8 -	max monthly precip
Variable 9 -	average monthly precip
Variable 10 -	min monthly precip
Variable 11 -	max 24 hour precip
Variable 12 -	average number of days w/thunderstorms
Variable 13 -	average wind speed
Variable 14 -	% frequency of observations w/vis < =2.5 mi
Variable 15 -	average % cloudiness
Variable 16 -	average station pressure

A measurement for Pearson's . strength of association between two variables was calculated. An absolute value of (r) indicated a strength of linear relationship between two variables, whereby (+1) equals a perfect positive relationship, (-1) equals a perfect negative relationship, and (0) equals no linear relationship at all. Any relationship between two climate parameters that revealed a combination of correlation coefficients greater than absolute value (0.3) and significance levels less than (0.05) were targeted to be retained for potential input into the clustering phase of the research. Table 2 is a listing of all variable combinations that fulfilled the established coefficient and significance level criteria.

Multi-collinearity of variables, meaning variables closely related (highly correlated) to at least one other variable, was very evident. The existence of multi-collinearity enabled the variables to be reduced to only those that were not statistically alike.

After reviewing the precipitation correlation data, TEC climatologists regarded the average monthly precipitation as the most revealing variable.[11] Consequently, variable(s) multi-collinear to average monthly precipitation were eliminated first. Variables not eliminated because of close correlation to average monthly precipitation were checked against one another. A final list of non multi-collinear precipitation variables was compiled and stratified by month.

After reviewing the temperature correlation data, TEC climatologists regarded the average monthly temperature as the most revealing variable. Variables multi-collinear to average monthly temperature were eliminated first. The average daily maximum and average daily minimum were considered to be second most important. If either of these temperature variables were not eliminated

[11] P.F. Krause and T. Niedringhaus, of TEC's Environmental Support Branch, provided expertise in the climatological discipline.

Table 2. Climate Variables with Correlation Coefficients Greater Than |(0.3)| and Significance Levels Less Than (0.05)

Precipitation

		Coef.	Sig.
January:	Max Monthly with Avg Monthly	.7582	.0000
February:	Max Monthly with Avg Monthly	.8448	.0000
March:	retain all variables	------	------
April:	retain all variables	------	------
May:	Max Monthly with Avg Monthly	.9105	.0000
June:	Max Monthly with Avg Monthly	.6021	.0003
	Max Monthly with Min Monthly	.5879	.0000
	Avg Monthly with Min Monthly	.6567	.0000
July:	Max Monthly with Avg Monthly	.9076	.0000
	Max Monthly with Min Monthly	.6746	.0000
	Max Monthly with Max 24 Hour	.5397	.0037
	Avg Monthly with Min Monthly	.8197	.0000
	Avg Monthly with Max 24 Hour	.5144	.0043
	Min Monthly with Max 24 Hour	.4072	.0350
August:	Max Monthly with Avg Monthly	.8885	.0000
	Max Monthly with Min Monthly	.6375	.0002
	Max Monthly with Max 24 Hour	.6011	.0009
	Avg Monthly with Min Monthly	.8220	.0000
	Avg Monthly with Max 24 Hour	.5747	.0011
	Min Monthly with Max 24 Hour	.5570	.0025
Sept:	Max Monthly with Avg Monthly	.8837	.0000
	Max Monthly with Min Monthly	.5127	.0038
	Max Monthly with Max 24 Hour	.6298	.0004
	Avg Monthly with Min Monthly	.6289	.0002
	Avg Monthly with Max 24 Hour	.7128	.0000
	Min Monthly with Max 24 Hour	.6017	.0009
October:	Max Monthly with Avg Monthly	.9003	.0000
	Max Monthly with Max 24 Hour	.6561	.0002
	Avg Monthly with Max 24 Hour	.8581	.0000
November:	Max Monthly with Avg Monthly	.8837	.0000
	Max Monthly with Max 24 Hour	.4798	.0113
	Avg Monthly with Min Monthly	.4941	.0055
	Avg Monthly with Max 24 Hour	.7250	.0000
December:	Max Monthly with Avg Monthly	.9152	.0000
	Max Monthly with Min Monthly	.5721	.0010
	Max Monthly with Max 24 Hour	.6218	.0005
	Avg Monthly with Min Monthly	.7410	.0000
	Avg Monthly with Max 24 Hour	.7805	.0000

Table 2 (continued). Climate Variables with Correlation Coefficients Greater Than |(0.3)| and Significance Levels Less Than (0.05)

Temperature

			Coef.	Sig.
January:	Abs Max	with Avg Day Max	.6587	.0000
	Abs Max	with Avg Day Min	.5412	.0002
	Abs Max	with Abs Min	-.3819	.0126
	Avg Day Max	with Avg Monthly	.5456	.0002
	Avg Day Max	with Avg Day Min	.6078	.0000
	Avg Monthly	with Avg Day Min	.7202	.0000
February:	Abs Max	with Avg Day Max	.3580	.0233
	Abs Max	with Avg Day Min	-.3293	.0380
	Avg Day Max	with Avg Monthly	.4418	.0043
	Avg Monthly	with Abs Min	.5292	.0004
	Avg Day Min	with Abs Min	.7334	.0000
March:	Avg Monthly	with Avg Day Min	.4116	.0126
April:	Abs Max	with Avg Day Max	.7967	.0000
	Abs Max	with Avg Monthly	.3480	.0239
	Abs Max	with Avg Day Min	.7988	.0000
	Avg Day Max	with Avg Monthly	.4689	.0017
	Avg Day Max	with Avg Day Min	.6745	.0000
	Avg Monthly	with Avg Day Min	.4438	.0032
	Avg Monthly	with Abs Min	.6330	.0000
May:	Abs Max	with Avg Day Min	-.4271	.0060
	Abs Max	with Abs Min	-.4052	.0095
	Avg Day Max	with Avg Monthly	-.4170	.0074
June:	Avg Day Max	with Avg Monthly	.5402	.0006
	Avg Day Max	with Avg Day Min	-.3512	.0328
	Avg Day Max	with Abs Min	-.3429	.0437
July:	Abs Max	with Avg Day Max	.9211	.0000
	Abs Max	with Avg Monthly	.8605	.0000
	Abs Max	with Avg Day Min	.7320	.0000
	Abs Max	with Abs Min	.5454	.0002
	Avg Day Max	with Avg Monthly	.9715	.0000
	Avg Day Max	with Avg Day Min	.8607	.0000
	Avg Day Max	with Abs Min	.7216	.0000
	Avg Monthly	with Avg Day Min	.9506	.0000
	Avg Monthly	with Abs Min	.7996	.0000
	Avg Day min	with Abs Min	.9297	.0000

Table 2 (continued). Climate Variables with Correlation Coefficients Greater Than |(0.3)| and Significance Levels Less Than (0.05)

			Coef.	Sig.
August:	Abs Max	with Avg Day Max	.9141	.0000
	Abs Max	with Avg Monthly	.8296	.0000
	Abs Max	with Avg Day Min	.7186	.0000
	Abs Max	with Abs Min	.4736	.0000
	Avg Day Max	with Avg Monthly	.9673	.0000
	Avg Day Max	with Avg Day Min	.8502	.0000
	Avg Day Max	with Abs Min	.6904	.0000
	Avg Monthly	with Avg Day Min	.9484	.0000
	Avg Monthly	with Abs Min	.7802	.0000
	Avg Day min	with Abs Min	.9155	.0000
Sept:	Abs Max	with Avg Day Max	.8865	.0000
	Abs Max	with Avg Monthly	.7684	.0000
	Abs Max	with Avg Day Min	.7018	.0000
	Abs Max	with Abs Min	.4115	.0000
	Avg Day Max	with Avg Monthly	.9629	.0000
	Avg Day Max	with Avg Day Min	.8235	.0000
	Avg Day Max	with Abs Min	.7516	.0000
	Avg Monthly	with Avg Day Min	.9378	.0000
	Avg Monthly	with Abs Min	.8207	.0000
	Avg Day min	with Abs Min	.8875	.0000
October:	Abs Max	with Avg Day Max	.8701	.0000
	Abs Max	with Avg Monthly	.7619	.0000
	Abs Max	with Avg Day Min	.6293	.0000
	Abs Max	with Abs Min	.3087	.0496
	Avg Day Max	with Avg Monthly	.9672	.0000
	Avg Day Max	with Avg Day Min	.8415	.0000
	Avg Day Max	with Abs Min	.6678	.0000
	Avg Monthly	with Avg Day Min	.9458	.0000
	Avg Monthly	with Abs Min	.7591	.0000
	Avg Day min	with Abs Min	.8621	.0000
November:	Abs Max	with Avg Day Max	.7590	.0000
	Abs Max	with Avg Monthly	.5718	.0001
	Abs Max	with Avg Day Min	.5221	.0007
	Avg Day Max	with Avg Monthly	.9675	.0000
	Avg Day Max	with Avg Day Min	.9097	.0000
	Avg Day Max	with Abs Min	.5687	.0002
	Avg Monthly	with Avg Day Min	.9777	.0000
	Avg Monthly	with Abs Min	.6443	.0000
	Avg Day min	with Abs Min	.6738	.0000

Table 2 (continued). Climate Variables with Correlation Coefficients Greater Than |(0.3)| and Significance Levels Less Than (0.05)

			Coef.	Sig.
December:	Abs Max	with Avg Day Max	.7872	.0000
	Abs Max	with Avg Monthly	.6235	.0000
	Abs Max	with Avg Day Min	.6155	.0000
	Avg Day Max	with Avg Monthly	.9755	.0000
	Avg Day Max	with Avg Day Min	.9285	.0000
	Avg Day Max	with Abs Min	.5006	.0012
	Avg Monthly	with Avg Day Min	.9807	.0000
	Avg Monthly	with Abs Min	.6145	.0000
	Avg Day min	with Abs Min	.6317	.0000

by 'average daily temperature', they were queued up as the next variable to be retained. A final list of non multi-collinear temperature variables was compiled and stratified by month.

The variables identified in Table 3 were selected from the original candidate list. They represent the critical variables to be included in a month by month statistical clustering process for the 43 climate stations. Added to this final list of variables were **'elevation'** and **'barometric pressure'**.

Table 3. Critical Variables Selected from Correlation Analysis for the Clustering Process

Month	**Precipitation**				**Temperature**				
	1	2	3	4	1	2	3	4	5
January	X	X	X			X	X		
February	X	X	X			X			X
March	X	X	X	X	X	X	X	X	
April	X	X	X	X		X			
May	X	X	X			X		X	X
June		X	X		X	X		X	X
July		X				X			
August		X				X			
September		X				X			
October	X	X			X	X			
November		X				X			
December		X				X			

Precipitation:
1 = Min Monthly
2 = Avg Monthly
3 = Max Monthly
4 = Max Monthly

Temperature:
1 = Absolute Max
2 = Avg Monthly
3 = Avg Daily Max
4 = Absolute Min
5 = Avg Daily Min

The months of July, August, September, November and December all use the same two variables--average monthly precipitation and average monthly temperature. Initially, this seemed unlikely, but TEC climatologists believe that the North Sea does have a moderating effect on Germany's weather and that it must do so on into late fall and early winter. Some German residents have confirmed that the harshest winter conditions do begin in January, which coincides with the month in which the variables do show change.

Clustering

Background. Clustering analysis algorithms attempt to imitate an otherwise subjective process of grouping observations into similar categories. "Within cluster analysis, little is known about the category structure. All that is usually available is a collection of observations whose category memberships are unknown. The objective is to discover category structures which fit the observations; in other words, find the natural groups. Clustering categorizes observations into groups such that the degree of natural association is high among members of the same group and low between different group members."[12]

Grouping the world into climates zones is not a new endeavor. Koppen compiled a schema of small-scale climate zones applicable to all countries of the world. This most widely recognized schema, however, is unacceptable for this research because entire countries cannot be categorized by a single climate. Larger scale schemas exist, but they lack any global continuity in presentation of scale or detail. Merging these larger scale schemas into one global digital presentation would have created an inadequate product. Developing statistically created climate zones with global continuity can be accomplished by using the clustering technique described herein. Clustering was used to pre-process the climate data off-line for input into the follow-on TIN phase of the research.

Wards Clustering Method. Data from the critical correlation variables previously selected from correlation analyses were input into a PC-based clustering software routine. Homogeneous groups of climate stations were then computed using the Wards clustering method with squared Euclidean distance to the cluster means calculated for each climate station. The distances were summed for all stations. At each step, the two clusters that merged together were those that had the smallest increase in the overall sum of the squared within cluster distances.[13] Wards method is an iterative, hierarchical process that depends on within group variance.

Cluster Group Selection. Determining if the number of clusters, or groups, generated by the cluster analysis was correct was a subjective process. Potentially, the number of groups was somewhere between 43 (one for every climate station) and 1 (all stations being grouped together as one). Several considerations went into determining the number of groups to select.

o **First Consideration: Dendrograms.** The initial consideration in determining the appropriate number of groups was to review hanging icicle tree graphs, or dendrograms, of the stations as they were shown to cluster together into homogeneous groups (see Figure A5, Example

---

[12] M.R. Anderberg, Cluster Analysis for Applications, Air Force Systems Command, United States Air Force, Academic Press, Inc., New York, New York, 1973, p. 2-4.

[13] NORUSIS/SPSS Inc., SPSS-X Introductory Statistics Guide, 1988.

Dendrogram). A subjective determination was made as to where a natural deviation occurred in the icicle-grouping process. The number of steps that occurred after this natural deviation on the dendrogram defined the number of cluster groups appropriate for grouping homogeneous climate stations.

o **Second Consideration: Coefficients.** A review was made of the statistical coefficients generated at each step of the clustering algorithm process. A natural 'break', or change, was identifiable in the coefficients whenever their value would markedly increase from a previously recognizable incremental pattern. Similar to dendrograms, the number of steps that occurred after a coefficient value break identified the number of clusters appropriate for grouping climate stations into representative climates. In Table 4 for example, three groups were selected because the break between 0.081 and 0.228 was viewed as the most significant change. Dendrogram icicle breaks and coefficient breaks were verified against each other to ensure they agreed with the number of cluster groups selected.

Table 4. Example of Clustering Coefficients

	Cycle	Group in	Group to Join	Coefficient
	.....	....	....	....
	......	....	....	....
	35	2	4	0.037
	36	7	20	0.044
	37	1	5	0.049
	38	7	11	0.080
	39	1	38	0.081
			BREAK POINT	
1)	40	7	37	0.228
2)	41	1	2	0.264
3)	42	1	7	0.514

o **Third Consideration: Grouping Rationale.** Once the number of cluster groups was determined for each month of the year, all stations that fall within those groups were identified. The TEC climatologists analyzed individual stations to see if the computer-derived groupings could be rationalized. The groupings were verified by visually inspecting the climate station data and comparing station groupings to elevation. For example, a very simplified clustering reliability check used was to ensure that a unique, high alpine climate station was identified as an isolated group during non-summer months. The station in question routinely showed up as a separate group unto itself. Any group(s) with one or two stations was closely examined and its minimal station memberships were rationalized.

o **Fourth Consideration: Geographic Mapping.** Climate station locations were plotted on a small-scale map to visualize how they spatially interrelated to one another and how they related to the oceans and highlands of the region (See Figure A6 - Geographic Map Of Germany Climate Stations). Stations that grouped together were evaluated to see if they were located in the same area. Isolated stations were examined and rationalized. In general, although coastal stations grouped together, as did the mountain stations, this was not the rule. Stations assigned to different cluster groups did not always group together when viewed in a two dimensional spatial orientation. This variability was hypothesized to be attributable in large part to the elevation differences between stations and to the orientation/aspect of the underlying local relief in the area. A review of the actual climate data validated the cluster grouping assignments because the values did appear slightly different wherever instances of spatially inter-related station locations occurred.

After the four-phase consideration process for cluster group designation was completed, final cluster groups were selected as the most reasonable (see Table 5.) Monthly climate station groupings were incorporated into the TIN phase of research.

Table 5. Final Cluster Groups by Month

Cluster Groupings: January
1 (A) 1-6 8-10 13 14 16-19 21-36 38-43
2 (B) 7 11 12 15 20
3 (C) 37

Cluster Groupings: February
1 (A) 1 2 4 9 10 16 18 19 23 25 28 29 31 38 42
2 (B) 3 5 6 8 12 13 14 17 21 22 24 26 27 30 32-36 39-41 43
3 (C) 7 20
4 (D) 11 15
5 (E) 37

Cluster Groupings: March
1 (A) 1-3 5 6 8 9 12-14 16 17 19 21-24 26 27 29-36 38-43
2 (B) 4 18 25 28
3 (C) 7 11 20
4 (D) 10 15
5 (E) 37

Cluster Groupings: April
1 (A) 1 2 7 23 28 29 31 42
2 (B) 3 5 6 8-10 12-14 16 17 19 21 22 24 27 30 32-36 38-41 43
3 (C) 4 18
4 (D) 11 15 20
5 (E) 37

Cluster Groupings: May
1 (A) 1-3 7 9 12 14 6 17 19 23 25 29 31 33 38 41-43
2 (B) 4 15 18 28
3 (C) 5 6 8 13 21 22 24 26 27 30 32 34-36 39 40
4 (D) 20 11
5 (E) 37

Table 5. (continued) Final Cluster Groups by Month

Cluster Groupings: June
1 (A) 1 6 30 36 38
2 (B) 2 7 12 15 16 20 28 311 42
3 (C) 3 5 8-10 13 14 17 19 21-27 29 32-35 39-41 43
4 (D) 4
5 (E) 37

Cluster Groupings: July
1 (A) 1 2 4 7 9 11 14-16 18 20 28 31 41
2 (B) 3 5 6 8 10 12 13 17 21-27 29 30 32-36 38-40 42 43
3 (C) 37

Cluster Groupings: August
1 (A) 1-3 5-10 12-17 19 21-36 38-43
2 (B) 4 11 18 20
3 (C) 37

Cluster Groupings: September
1 (A) 1-3 5 6 8-10 12-14 16 17 19 21-36 38-43
2 (B) 4 7 11 15 18 20 37

Cluster Groupings: October
1 (A) 1 5 6 8 10 13 14 17 19 21 22 24 26 27 32-34 38-41 43
2 (B) 2-4 9 12 16 18 23 25 28-31 35 36 42
3 (C) 3

Cluster Groupings: November
1 (A) 1-6 8-10 12-14 16-19 21-36 38-43
2 (B) 7 11 15 20
3 (C) 37

Cluster Groupings: December
1 (A) 1 5 6 8 10 13 14 21 22 24 26 32 34 38 40 41 43
2 (B) 2-4 9 12 16-20 23 25 27-31 33 35 36 39 42
3 (C) 7 11 15
4 (D) 37

Selection of a Point-in-Polygon Method

A point-in-polygon method was used to determine the closest climate station to a user-defined observer location. Three differing methods for polygonizing the area around the climate station were considered: (1) Thiessen polygons, (2) Manual-derived polygons, and (3) Triangulated irregular network polygons.

Thiessen polygons. Thiessen polygons offered what seemed to be a viable automated method for generating polygons around individual climate stations, with the size of the polygons proportionally related to the distance between adjacent stations. A user-defined observer location would be mapped to see which polygon it would fall into. The observer location would then emulate the climate parameters of the climate station occupying that polygon. A problem with this technique was that it did not provide knowledge about the surrounding polygons. This information, if available, could have indicated when an observer location was actually adjacent to a polygon containing a station with similar elevation. With thiessen polygons, there was only one option for station selection. The thiessen polygon method was not regarded as the best option.

Manually derived polygons. A second option for mapping the cluster data was to manually determine polygons by interpolating halfway between different cluster designated groups and then to convert them into digital overlays (see Figure 1). Manually contouring and then digitizing point data was considered laborious and subjective. However, the interpolation process could not be automated because the climate stations themselves retained no real numeric values, only climate grouping assessments. There was only one option for a climate station selection, but observer locations were conceivably located between two or more climate zones.

Triangulated irregular network polygons. A third method for developing climate station polygons was the TIN method. Typically, TIN's are associated with vertical elevation data. They provide an alternative to portraying elevation data in a gridded uniform manner. A TIN shows vertices at strategically selected highs and lows across an area of terrain. This selection process is designed to minimize the collection and storage of elevation points to only those points that represent significant changes in relief. For example, areas of rugged terrain are represented by a denser pattern of TIN vertices than are areas of flat terrain.

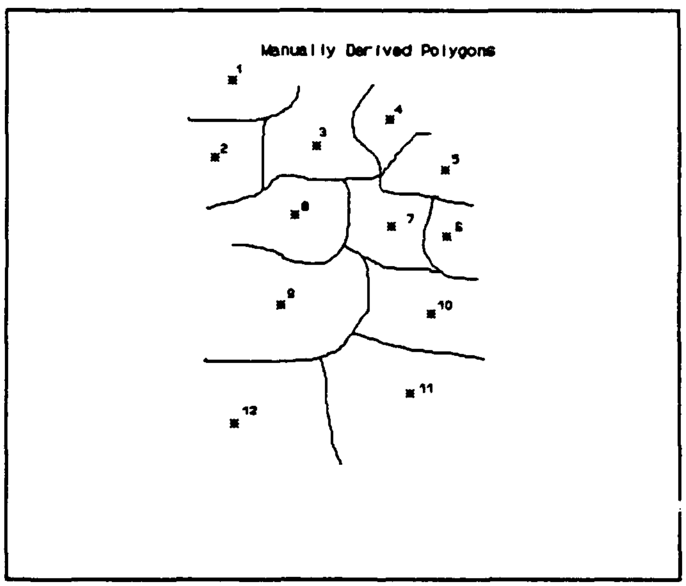

Figure 1. Manually Derived Polygons.

In this research, the TIN vertices were not representing elevation points but rather climate stations attributed with the respective cluster groups to which they had been previously assigned. The TIN polygons always have three climate station vertices, or nodes, which increases the number of stations to choose from three-fold. Twelve TIN's were generated for the Germany data set; one for each month of the year. Cluster groups associated with each station changed on a monthly basis. Attribute information pertinent to each TIN, such as cluster group, was stored in a retrievable data base. Cluster group information was relied on to determine if an observer location was between climate zones.

A TIN can be generated and displayed in many forms. This research used the Delaunay TIN, which is one of the simpler, intuitive types of TIN's. Delaunay TIN's are unique because an imaginary circle can be drawn surrounding any three triangle vertices, which does not incorporate any additional triangle vertice(s) (see Figure 2).

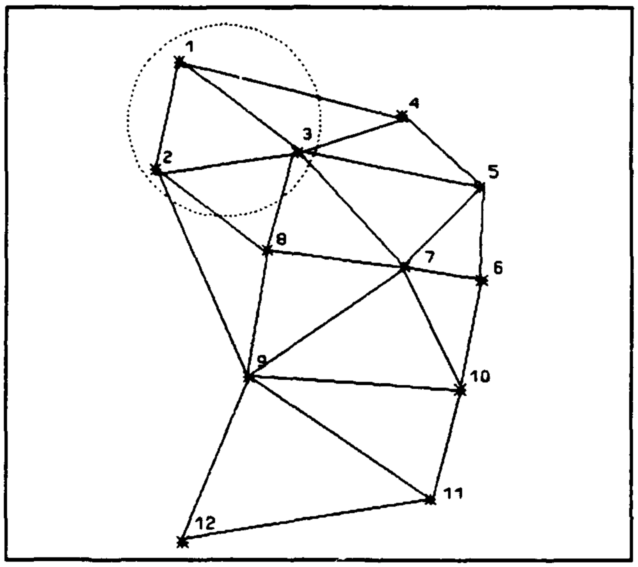

Figure 2. Delaunay TIN.

Reliability Logic. The purpose behind the reliability logic was to alert the user to inconsistencies exhibited among surrounding climate station data. Several options exist to determine the reliability of climate data passed to the computer model. In general, the greater the number of climate zones surrounding an observer location, the less reliable the data.

Option 1: When a climate station was selected from the primary TIN vertices and when all three of these vertices had the same cluster group designator, the observer location was considered to be centered within a relatively homogeneous area (or micro-climate). For example, a primary TIN with Node 7 is designated Cluster B; Node 9 is designated Cluster B; and Node 10 is designated Cluster B as an example of Option 1.

Option 2: When a climate station was selected from the primary TIN vertices and when two stations are alike with one different, it is interpreted that the observer is between two different climate zones. For example, Node 7 is designated Cluster B, Node 10 is designated Cluster B, and Node 9 is designated Cluster C (see Figure 3). The TIN triangles of Option 2 type were observed in one-third of the study sites.

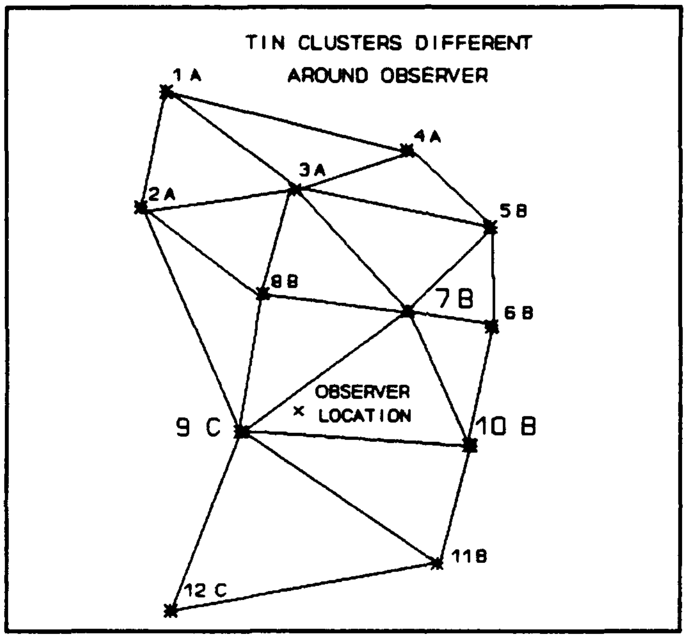

Figure 3. Two Surrounding Station Nodes Alike, One Different.

Option 3: When a climate station was selected from the primary TIN vertices and when all three stations were defined as being from different clusters, the observer location is interpreted to be in a "grey" area somewhere between three climate zones. For example, Node 7 is designated Cluster A, Node 9 is designated Cluster B, and Node 10 is designated as Cluster C.

Option 4: When a climate station was selected from secondary TIN vertices, the opportunity is introduced to return a message to the user of a maximum possibility of four different climate zones--one for each primary vertice and one for the secondary TIN vertice selected. This occurred once during the 30 observations that were tested and evaluated.

Station Selection Process. The TIN routine passed back to the user the names of six climate stations that surrounded the observer location. Three of the six stations were the vertices to the triangle immediately surrounding the observer (nodes 7, 9, and 10 from Figure 3), while the remaining three stations identified secondary vertices to the three adjacent triangles (nodes 6, 8, and

11). Ratio information for each station was also passed back with the station names. Ratios were useful to determine proximity to the observer location because they related to distance from a station to the observer location. Once the six stations and their respective proximity were determined from the TIN routine, the station with the highest ratio was always the first selected for consideration.

Once a station was selected for consideration, its elevation value was compared against the elevation value for the defined observer location. The difference between station and observer was not allowed to exceed an absolute value threshold of 500 feet[14] (+/- 500'). If the 500-foot difference was exceeded for the station, it was inferred that the observer location was more closely related to an alternative station. Elevation variables always took priority over the proximity variable. The next closest station was consequently examined for selection. If none of the three station vertices of the primary TIN triangle was within 500 feet of the observer location elevation, the observer location was interpreted to be at a point unique to the surrounding geographic area. In effect, none of the three primary stations could be used to mimic the climate parameters.

For example, refer to Figure 4. A climate station (call it 'A') is closest to a user-defined location on the ground. However, the station and the user location are vastly different in terms of vertical elevation (1900 versus 700 feet). This vertical difference makes it very difficult to justify that Station A has a climate similar to the observer location. In all likelihood, it does not. The model user has no convenient way of knowing when climate data were selected from a station that was not very similar in weather pattern to the observer location. Users are left to assume that the climate information is reliable. A second station (call it 'B') is just slightly farther away from the user location than the first station (27 miles versus 20 miles) and, therefore, overlooked by the computer selection. The elevation of Station B, however, closely matches the user location and would have been a better station selection to emulate the climate of the observer location.

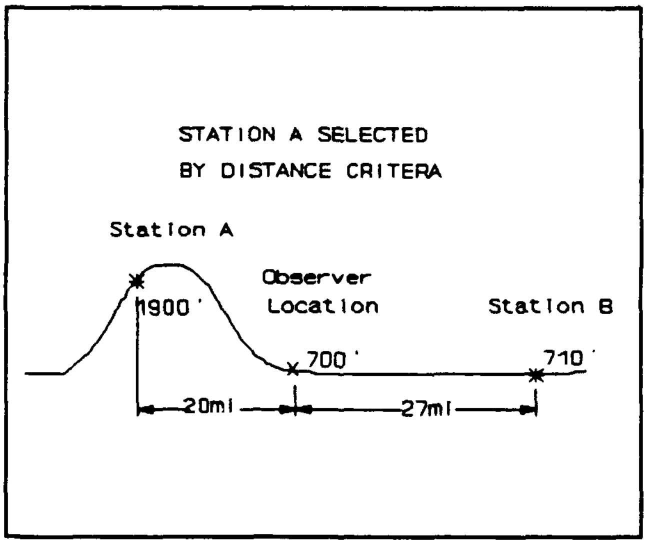

Figure 4: Distance Versus Elevation Example

[14] The 500 foot threshold was a value subjectively arrived at following careful consideration by USATEC climatologists.

Whenever none of the primary TIN vertices had an elevation that matched the 500-foot tolerance of the observer location elevation, a routine to choose one of the three secondary triangle vertices was invoked. Selecting the secondary vertice was based first on proximity, followed again by a check to verify that the elevations between station and observer location were within the 500-foot tolerance. This iterative process continued until an elevation match was made.

If no stations met the elevation criteria, the initial station chosen but later eliminated because of elevation difference, was selected. A warning message was issued to the user declaring that the climate data originated from a station that was not necessarily representative of the user-selected observer location.

Once the proximity and elevation criteria established the appropriate climate station to select, the model examined the climate station cluster group designation and returned the correct "Option" message to the user. These messages alerted the model user of competing climate zones. This supplemental information was not intended for actual input into later computer modeling but rather as a measure of reliability for the user to reference.

Although adjustments could have been made, they were not used for climate parameter "average monthly temperature" to account for the elevation difference between observer location and station elevation. A normal lapse rate adjustment could have easily been computed to record a world wide average decrease of 3.5 degrees Farenheit for every increase of 1000 feet of vertical elevation.[15] However, a maximum of 1.75 degrees of temperature difference, based on a maximum 500-foot elevation threshold, was not deemed worthwhile at this time. The lapse rate adjustment may be included in future revisions to this methodology.

Example "Blackbox" Search Query: The following example illustrates this newly devised methodology for climate station selection. Given a known observer location, identify the weather station most likely to retain comparable climatic conditions.

1. Find an observer location's geographic relationship to all climate stations by a point-in-polygon method. The TIN routine is invoked and identifies the six surrounding climate station nodes of the TIN triangles.

---

[15] Temperature Conversion Algorithm:

OE - SE = X
X / 1000' = (%)
% * 3.5 = (C)
SAMT + (C) = <u>Observer location temperature</u>

Given a normal lapse rate adjustment of 3.5 degrees F for every increase in elevation of 1000 feet:

OE =Observer Elevation
SE =Station Elevation
X =Difference in Elevation between OE and SE
(%) =Percentage of Lapse Rate Change
(C) =Amount of Lapse Rate Change
SAMT =Station Average Monthly Temperature

2. The TIN reports that the observer location is surrounded by climate station primary nodes with respective cluster groups of A, A, and B. This implies the observer location is in an area of uncertainty between stations from differently clustered climate groups.

3. The nearest climate station node to the observer location, as defined from the TIN ratios, has an elevation of 598 feet and a cluster group designation of (A). The observer location elevation is 65 feet (598 - 65 = 533 feet). The 533 feet is not within the established 500-foot absolute value threshold; so the station is not accepted.

4. The next closest primary TIN vertice is examined and it has an elevation of 298 feet and a cluster group designation of (B) (598 - 298 = 300 feet). The 300 feet is well within the 500-foot threshold. Therefore, the station is accepted as the most reasonable in similarity to the observer location. Climate information is provided for the observer location from this climate station. The climate station found at cluster group (A) is also passed to the user as supplemental information. A warning is issued to the user that the observer location is between zones.

Test and Evaluation

Thirty observer location points within an area in Germany bounded by (48N, 7E) and (54N, 14E) were examined to determine which climate stations would be selected using the distance only method and which stations would be selected using the new methodology that places emphasis on an elevation variable. Within the 6 by 7 degree geographic area, there were 43 climate stations. The month chosen for testing was February because it contained five different climate cluster groups, the maximum amount defined. Twenty of the 30 observer locations tested returned identical climate stations regardless of methodology chosen for station selection. Ten stations, or 33 percent, did not duplicate the climate station selection. Elevation was the critical variable in each of the differing station selections. Results of the testing are found in Figure A7, Test and Evaluation Data.

ANALYSIS

The comparison of differently selected climate stations, in terms of elevation, was revealing. The data for each station were different, naturally, and the consequences of using the wrong data in a model became more apparent. Mobility modeling, for example, depends on precipitation data to determine soil moisture strength. Geographic observer location #24, found under heading Coordinate in Figure A7, is located at 344 feet elevation. This observer location returned the Clausthal-Zellerfeld station at 1919 feet elevation using the distance only method and returned the Wittenberge station at 85 feet elevation using the new method. Precipitation data differences between the two stations were significant. For example, average monthly precipitation for February shows Clausthal-Zellerfeld with 4.2 inches and Wittenberge with 1.2 inches. This difference would alter the outcome of a soil moisture strength analysis, which in turn would alter the output derived from a cross-country mobility model.

Geographic observer location #1 returned different climate stations with vastly different elevations. The old algorithm returned Feldberg station at 4908 feet, while the new algorithm returned Zurich station at 1617 feet. A normal standard lapse rate temperature adjustment measured against each 1000 feet of elevation change suggests that the temperature for Feldberg should be, at a minimum, approximately 10 to 11 degrees cooler than those of Zurich. A temperature change of that magnitude might be enough to erroneously effect a computer model. For example, mobility GO areas may be identified across ground determined to be frozen based on the temperature parameters provided. These "frozen" grounds may in fact be muddy, impassable tracts of land given a 10-degree rise in temperature.

Of the 30 observer locations tested, there was strong support for providing supplemental information to the user regarding relationship of the observer location to the surrounding climate zones. Three cases existed where no climate zones appeared to be totally reliable due to the 500-foot elevation threshold not being met by either primary or secondary stations. Ten cases existed where the observer location was identified as positioned between two climate zones. Sixteen cases existed where the observer location was identified as positioned between three climate zones. One case existed where the observer location was identified as positioned between four climate zones because a secondary TIN climate station was selected as most appropriate and each of the primary TIN vertices had unique climate group designations. Despite a relatively dense network of climate stations found within Germany, the five different February climate zones defined previously by the clustering program resulted in the observer location never being positioned within one single climate group. However, had the test been examined against the three cluster groups from July, for example, it is anticipated that observer locations positioned within single climate zones would have frequently occurred.

CONCLUSIONS

The processes of correlation and clustering were successfully demonstrated across a small region and could be replicated to the entire worldwide BEES climatic data base. The TIN process worked successfully across the entire climatic data base. A 33 percent improvement from the previous method of selecting the closest climate station without regard for terrain, as compared with the new method that used elevation values, was recognized. Climatic data input was more reliable with the new elevation method, which means computer model output was more accurate.

A better understanding of the reliability of the climate data was also reached by providing the cluster group for each TIN climate station node. Whenever a location resided between climate zones, a message alerted the user to the situation and implied a degree of uncertainty about the data. Observer locations that are between climate zones will not be an exception to the norm as evidenced by the 30 observation locations tested in this research. Until such time as computer model users can regard the selection of a climate station and its corresponding data as completely reliable, notifying the user of variable climate zones appears worthwhile.

The inclusion of an elevation variable improved the selection process, although additional variables need to be added. Proximity to water bodies should be the next variable entered into the program. As new variables are entered into the TIN program and the station selection is more accurately defined, the between climate zones supplemental information may be eliminated. The methodology described in this report can be generalized to include the interests of meteorological station data selection and is not limited to just climate data. Point data could be correlated, clustered, and manipulated within a TIN for either type of data.

Regarding the clustering techniques used, the number of different climate groups defined was greatest in February through June (five groups each) and least from July through November (three groups each). This distinct difference in the number of climate groupings suggests that Germany's climate patterns are most variable, and therefore most difficult to emulate, during the winter and spring months. Warmer months have far less contrast between air masses than do the colder months. In the colder months, solar radiation is less prominent, which then emphasizes the differences between the resident air masses.

Depending on the climate parameters desired to be emulated from a climate station to an observer location (i.e. temperature, precipitation, humidity), a particular combination of independent variables could be used to best determine the correct climate station to access. A weighting and ranking look-up table scheme of independent variables could be developed which best addresses each of the climate parameters. The independent variables would never change, but their inclusion in the selection process could be readily modified via the weighting criteria. Pre-determined combinations of variables, with their respective weights subjectively determined for now by climatologists' expert opinion, would be compiled for each climate parameter. Variables critical to solving a climate parameter would be assigned greatest weight while insignificant variables would be assigned low or null weights.

APPENDIX

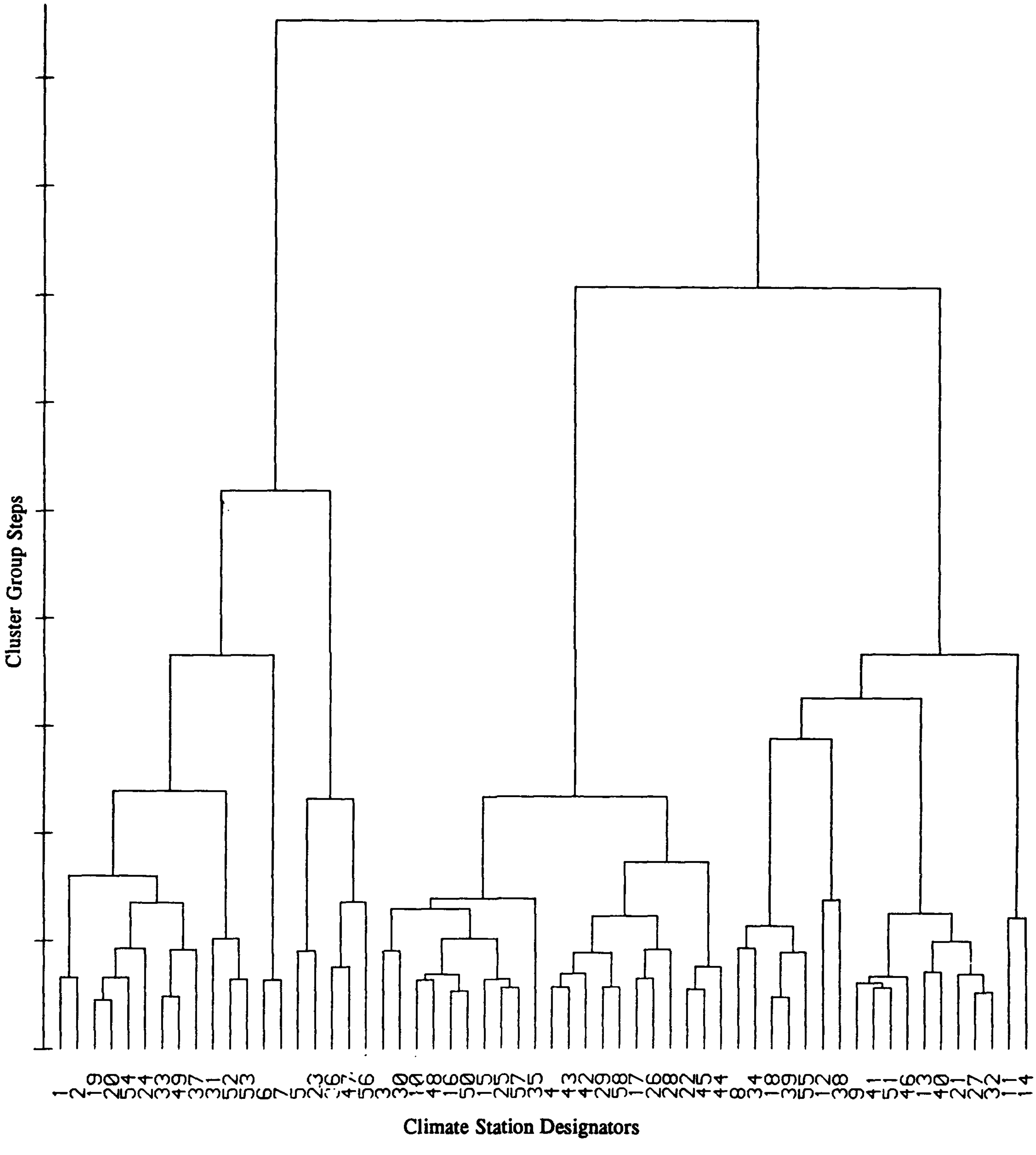

Figure A5. Example Dendrogram.

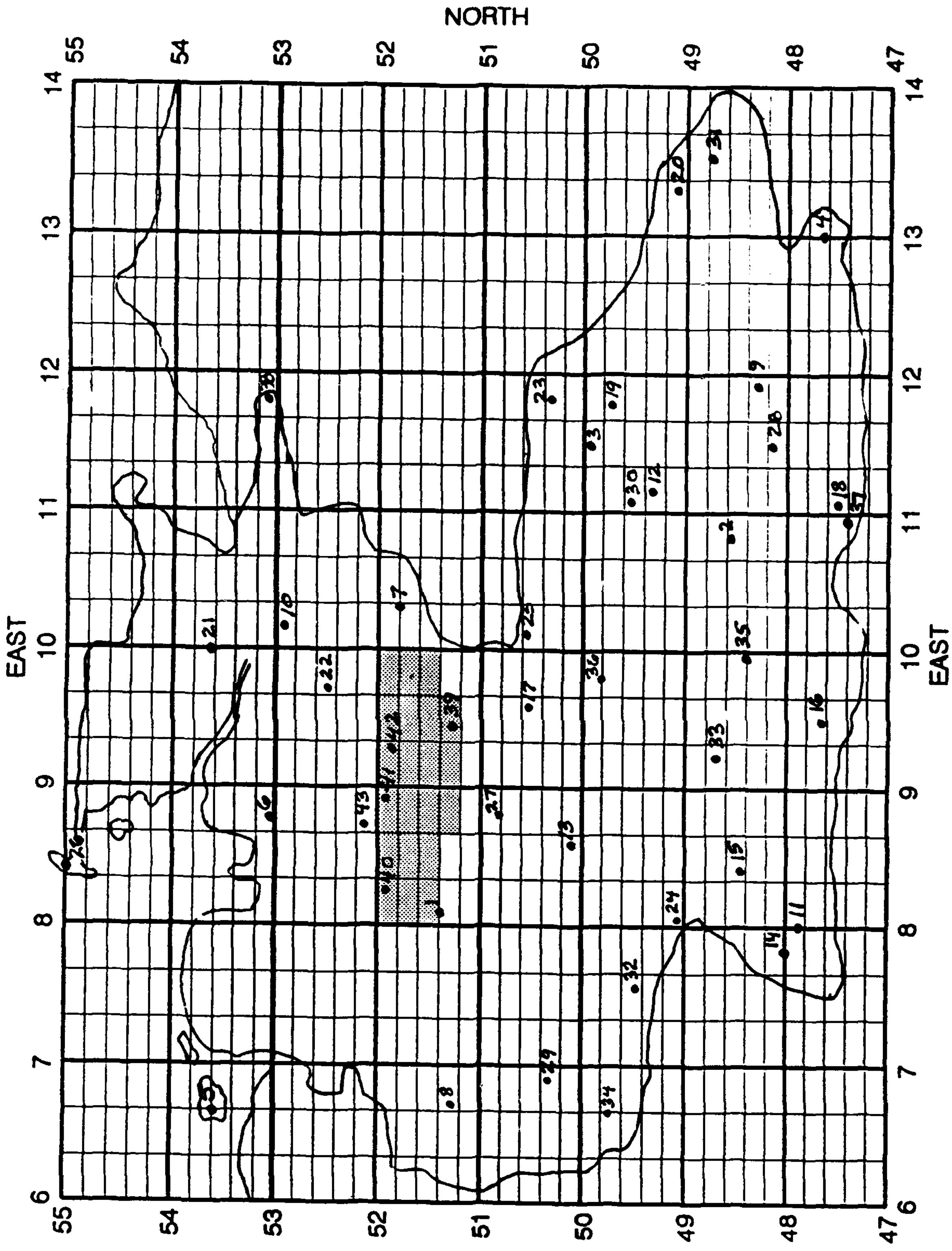

Figure A6. Geographic Map of German Climate Stations.

TEST AND EVALUTION DATA			
#	Coordinate	Closest Station	TIN Method
1	48N, 8E	Feldberg	Zurich
2	48N, 9E	Friedrichshafen	Frendenstadt
3	48N, 10E	Ulm	Ulm
4	48N, 11E	Augsberg	Garmisch
5	48N, 12E	Augsberg	Augsberg
6	49N, 8E	Karlsrhue	Karlsruhe
7	49N, 9E	Stutgart	Stutgart
8	49N, 10E	Ulm	Ulm
9	49N, 11E	Augsberg	Augsberg
10	49N, 12E	Augsberg	Augsberg
11	50N, 8E	Nuerburg	Frankfurt
12	50N, 9E	Frankfurt	Frankfurt
13	50N, 10E	Wurzburg	Wurzburg
14	50N, 11E	Bayreuth	Bayreuth
15	50N, 12E	Bayreuth	Bayreuth
16	51N, 8E	Arnsberg	Marburg
17	51N, 9E	Marburg	Marburg
18	51N, 10E	Fulda	Fulda
19	51N, 11E	Kaltennordheim	Fulda
20	51N, 12E	Hof	Fulda
21	52N, 8E	Arnsberg	Bremen
22	52N, 9E	Hannover	Hannover
23	52N, 10E	Clausthal-Zellerfeld	Hannover
24	52N, 11E	Clausthal-Zellerfeld	Wittenberge
25	52N, 12E	Wittenberge	Wittenberge
26	53N, 8E	Bremen	Bremen
27	53N, 9E	Bremen	Bremen
28	53N, 10E	Fassburg	Fassburg
29	53N, 11E	Wittenberge	Wittenberge
30	53N, 12E	Wittenberge	Wittenberge

Figure A7. Test and Evaluation Data.